05

Executando o teste de normalidade

Transcrição

[0:00] Beleza, pessoal? Vamos agora começar realmente a trabalhar com testes de hipóteses.

[0:05] A gente, no curso anterior, a gente já começou a mexer com inferência estatística, isso aqui é uma continuidade desse assunto.

[0:11] Lá a gente viu estimação, estimação pontual, estimação intervalar. Aprendemos a calcular o intervalo de confiança, tamanho de amostra.

[0:20] A gente aprendeu a obter probabilidades de uma distribuição normal, o que era nível de confiança, nível de significância. Todas essas coisas que a gente vem aprendendo lá, a gente vai aplicar agora dentro de testes de hipóteses.

[0:32] A gente vai ver aplicação prática de tudo isso, em conjunto, aqui dentro de teste de hipóteses. Legal?

[0:38] O que é teste de hipóteses? É uma regra de decisão que ajuda a gente a avaliar hipóteses feitas sobre os parâmetros populacionais.

[0:49] E aceitar ou rejeitar elas como provavelmente verdadeiras ou falsas, tudo isso tendo como base uma amostra.

[0:56] É o que a gente já começou a ver. A partir de uma amostra a gente tem uma estimativa de um parâmetro da população e vê se ele é representativo ou não.

[1:03] Uma coisa legal aqui é que a gente pode testar várias coisas. O primeiro teste aqui, que eu vou dar um susto em vocês como eu prometi no vídeo anterior, é o teste de normalidade.

[1:10] Antes de a gente ficava vendo, falava assim: essa variável é normal, não é normal? Como a gente fazia isso? Vendo se nela aparecia um sininho, só visualmente.

[1:19] Mas existe uma forma de a gente fazer isso mais formalmente, um teste mais robusto, onde ele tenha um valor estatístico, diga-se, com tanta probabilidade que você pode aceitar ou rejeitar a hipótese dessa distribuição, se seguiu a normal. Legal?

[1:36] Esse cara a gente vai ver também aqui. Outros exemplos de porque a gente está usando teste de hipóteses, um exemplo bem clássico, é aquela coisa da diferença de renda entre sexos.

[1:46] Homens ganham mais, mulheres ganham menos. Será que isso realmente é verdade, com base em um dado concreto que a gente tenha?

[1:55] Por exemplo, o nosso Dataset aqui tem informação disso, a gente vai testar isso no nosso treinamento. Uma outra coisa, por exemplo, também é.

[2:01] Um fabricante te dá uma informação e você tem que tomar aquilo como verdade. A gente pode testar se essa afirmação do fabricante é verdadeira ou não usando o teste de hipóteses.

[2:13] Então, vamos lá. Vamos começar com o teste de normalidade. Como eu disse, eu vou dar um susto em vocês - vai ter alguns conceitos que você não vai entender mesmo.

[2:21] Mas a ideia realmente é essa, só para você ver como faz, aplicar um primeiro teste. Depois, nos próximos vídeos, a gente vai fazer esse passo a passo, e eu vou mostrando para você como calcular essas estatísticas na mão, com o lápis. Ok?

[2:35] E no final eu mostro para você como você executa um teste como eu vou fazer agora, pegando uma funcionalidade do Scipy, do Statsmodels - do Python - para executar isso rapidamente.

[2:46] Então, vamos lá. Primeira coisa, eu vou importar esse cara aqui, que é o Normaltest. Eu vou usar esse, é o mais simples, é bem simples, é um teste de normalidade.

[2:53] Está aqui já a documentação dele. Uma coisa que eu quero chamar a atenção: é sempre bom ler a documentação desses testes que a gente está utilizando.

[2:59] Por quê? Lá ele te diz qual é a hipótese nula do teste. A gente já vai entender o que é isso, hipótese nula.

[3:05] O que ele está testando exatamente, ele vai dizer para você aqui.

[3:09] Nesse caso aqui, ele está dizendo, e eu já deixei até aqui, que a normaltest testa a hipótese nula, que é o H0, vamos conhecer, de que a amostra é proveniente de uma distribuição normal.

[3:22] Essa é a hipótese nula. É isso que eu tenho que rejeitar ou aceitar de acordo com o que a gente vai obter de resultado do teste.

[3:30] Vamos importar. Aqui eu já deixei pronto. From Scipy.Stats Import Normaltest. Rodou isso. Perfeito.

[3:44] O que eu quero fazer aqui, eu vou definir um nível de significância. Eu vou colocar isso dentro de uma variável. A gente já conhece o que é significância, o que é nível de confiança.

[3:56] O nível de confiança, trocando em miúdos, é a probabilidade de dar certo, do meu estimador estar correto; e o de significância é dele estar errado. Perfeito?

[4:10] Eu tenho uma célula aqui vazia, depois eu apago. Significância, vou colocar dentro desta variável, eu vou dar 5%, 0,05. É aquela significância padrão, nível de confiança 95%.

[4:27] Testando a variável renda do nosso Dataset, que a gente abriu no vídeo anterior. Vamos pegar a renda. Vamos visualizar primeiro como é que essa renda está se comportando.

[4:36] Então dados.Renda.hist. O Pandas tem a função que já é herdada do Matplotlib para a gente plotar um histograma.

[4:46] Eu vou botar aqui dentro um parâmetro que se chama Bins e eu vou falar 50. Ele está dizendo que eu quero 50 barrinhas no meu histograma. Vai ficar um pouco mais fácil de visualizar.

[4:56] Então está aqui. Visualmente, aquele nosso padrão de visualizar teste e visualizar uma normal, eu esperaria que a normal tivesse uma forma de sino, e eu estou vendo aqui que não tem esse padrão.

[5:13] Parece que é uma coisa assim. Sobe, desce e vem para cá. É uma coisa meio assimétrica. A gente já viu isso nos outros cursos.

[5:27] Visualmente eu consigo perceber que isso não segue uma normal, mas vamos testar isso usando um teste formal, para a gente ter uma estatística e dizer assim.

[5:37] Não, realmente essa variável não segue uma distribuição normal. Com certeza. Sem precisar visualizar.

[5:46] Vamos começar. Vou fazer direto o teste. É só chamar, Normaltest, que a gente já importou lá em cima, e passar o dado para ele. Dados.Renda. Roda isso. Ele já rodou o teste, está aqui.

[6:03] O Output desse teste são dois valores. Um é o Estatística de teste e o outro é o P valor. Tudo isso a gente vai conhecer ao longo do nosso curso.

[6:16] Como ele tem dois Outputs, é uma dupla, eu posso passar isso aqui para duas variáveis. Eu vou chamar de Stat_test, vírgula, P valor. Perfeito. E eu posso printar esses caras aqui em baixo.

[6:41] Posso separar eles porque eu posso querer, precisar usar isso para alguma outra coisa. A gente vai utilizar o P valor aqui.

[6:52] P valor. Copiar aqui e colar aqui embaixo. E a gente consegue visualizar eles separadamente.

[6:59] Aqui em cima eu já deixei uma regrinha de decisão. Quando é que eu rejeito o H0? É o que o H0? A gente já viu aqui em cima. O H0 é a afirmação de que a amostra é proveniente de uma distribuição normal.

[7:15] A regra de rejeição é simples: a minha variável de P valor é menor ou igual a 0,05. Que é o quê? O Alfa. Que é o quê? A minha significância, aquela variavelzinha que eu criei lá em cima.

[7:41] O que ele está dizendo? Que é verdadeira essa afirmação aqui. Então, o que eu faço? Eu rejeito o H0. Que é o quê, novamente? É a hipótese de que a amostra é proveniente de uma distribuição normal.

[7:54] O que ele está dizendo aqui é que não, não é proveniente de uma distribuição normal, eu rejeito essa hipótese aqui totalmente. Legal?

[8:01] Visualmente a gente confirma isso e agora, com uma estatística mais formal, a gente também confirma isso.

[8:08] Agora, vamos ver uma variável que realmente siga uma distribuição normal e vamos ver o que esse teste mostra para a gente, só para a gente sentir isso.

[8:14] A variável Altura, como eu disse no vídeo anterior e venho falando isso nos outros cursos, fui eu que gerei ela a partir de, justamente, uma distribuição normal, que é uma variável aleatória proveniente de uma normal.

[8:28] Vamos visualizar de novo, só para a gente ter certeza de tudo que eu estou falando. Altura.hist. E vou botar o mesmo bins, igual a 50, para ficar igualzinho.

[8:42] Olha lá que beleza, o sininho. Bonitinho aqui, certinho, com média em 1 metro e 70. Eu criei assim mesmo, de propósito.

[8:52] Aqui, novamente, a regra de rejeição. Eu posso copiar aqui esse cara. Vamos copiar, para a gente adiantar o nosso lado. E já vamos copiar esse daqui também, é o mesmo cara.

[9:05] Nós vamos substituir o P valor e o Stats aqui. A única coisa que eu tenho que fazer aqui é mudar de Renda para Altura e rodar o teste. Olha lá. A estatística está aqui em cima.

[9:19] Depois a gente vai entender o que é estatística. Agora a gente não está usando, mas a gente vai utilizar ela para tirar algumas decisões. E o nosso P valor, que é o principal. Todo teste que você vai rodar vai plotar o P valor para você, não se preocupe.

[9:33] E a regra de decisão é sempre essa. Olha aqui, já é um valor altinho, quase um. Vou rodar aqui a minha regrinha de decisão, que eu criei aqui.

[9:44] E o que ele está dizendo? Falso. Eu rejeito H0 se valor de P menor ou igual 0,5. O que ele está falando é que não, que é falso. Isso daqui é falso.

[9:54] O que eu faço? Não posso rejeitar. Eu não posso rejeitar a hipótese nula, H0, de que a amostra é proveniente de uma distribuição normal. E realmente é - fui eu que criei.

[10:07] E está aqui. Visualmente, realmente parece uma normal. Bonitinho o sino.

[10:12] Pessoal, é isso que eu queria mostrar nesse vídeo, esse susto inicial. Não se preocupem, a gente vai ver o que é P valor, a gente vai aprender a calcular um P valor na mão, na marra.

[10:21] H0, hipótese nula, rejeição, aceitar hipótese alternativa. Os níveis de significância a gente já conhece. Mas todas essas coisinhas que eu fui falando aqui, a gente vai ver passo a passo.

[10:33] No próximo vídeo a gente já vai entender os passos. E no outro a gente vai começar a botar a mão na massa. Vamos fazer testes, passo a passo, e depois no final, eu mostro como é que faz isso, de uma forma bem simples e prática, utilizando funcionalidades do Python. Beleza?

[10:46] No próximo vídeo a gente vê as etapas básicas de um teste. Até lá.